The Healthcare Administrator's Associate: An Experiment in Distributed Healthcare Information Systems

Jerry Fowler and Gale Martin
Microelectronics and Computer Technology Corporation (MCC)
Austin, Texas
{jfowler,gmartin}@mcc.com

The Healthcare Administrator's Associate is a collection of portable tools designed to support analysis of data retrieved via the Internet from diverse distributed healthcare information systems by means of the InfoSleuth system of distributed software agents. Development of these tools is part of an effort to enhance access to diverse and geographically distributed healthcare data in order to improve the basis upon which administrative and clinical decisions are made.

INTRODUCTION

As more and more healthcare institutions recognize the value of the Internet for communications, the possibility of wide–area electronic access to healthcare information grows. This can lead not only to improvement of statistical analyses by increasing the size of the samples analyzed, but to improved clinical care through access to a patient's longitudinal medical record as represented in components whose geographic dispersal at various clinics reflects the patient's diverse attempts to obtain medical care.

While public awareness of the power of the Internet grows, a consensus is emerging that the healthcare industry must address healthcare consumers' twin perceptions of increased medical costs and decreased quality of care. Moreover, healthcare providers face the impossible task of mastering a huge and rapidly growing body of medical knowledge. The industry is turning toward data-driven approaches to meet these challenges, by developing more automated ways of measuring excellence. Increasing the quality, quantity, and detail of data collected about patients and providers; and increasing the ability of providers and consumers to interpret these data through user friendly data analysis and decision support tools, can assist in this process. The goal is to provide feedback to the healthcare system to promote effective, efficient healthcare delivery.

MCC is contributing to this national effort through the collaboration of its InfoSleuth project with the Healthcare Open Systems Trials (HOST) consortium. The InfoSleuth project is currently developing an underlying information infrastructure for integrated access to distributed, heterogeneous databases, as well as data mining tools for making use of these integrated data for developing flexible outcome-based evaluation capabilities.

This paper describes the Healthcare Administrator's Associate (HAA), which is a prototype of a data analysis and decision support tool supported by InfoSleuth's distributed agent infrastructure for management of data reposing in physically or logically separated heterogeneous resources.

HEALTHCARE DATA ANALYSIS

Evidence of the development of data-driven metrics of quality can be found in public and private sectors, and at multiple levels of the healthcare industry. The Agency for Healthcare Policy and Research of the U.S. Department of Health and Human Services is working towards the standardization of hospital in-patient data through the establishment of a huge, nationwide database contributed by 17 states with many of these contributions dating back to 1988. Similarly, the U.S. National Institute of Standards and Technology (NIST) is funding advanced research and development efforts to develop technologies for integrating heterogeneous healthcare databases. The Joint Commission on Accreditation of Healthcare Organizations is encouraging private compliance by healthcare facilities with the requirement that, by the year 1999, a prerequisite for hospital accreditation will be the adoption of an approved system for data collection and data–driven quality measurement. Similar trends can be seen from

1091—8280/97/$5.00

the employers who provide managed care plans for their employees, in the development of the National Committee for Quality Assurance HEDIS system for comparing the quality of these plans.

These efforts are leading to the development of two types of complementary data-driven approaches: (1) outcomes-based evaluation of providers, for use primarily by administrators and consumers; and (2) evidence-based practice guidelines, for use by clinicians in determining the best treatment practices for individual patients. Currently, there are considerably more data available for outcomes-based quality metrics, which are developed using existing standardized schemes for coding patient diagnoses and procedures, such as ICD9 and DRG codes, and more diverse coding schemes for recording patient demographic and cost of care data.

Evidence-based practice guidelines require considerably more detailed data about patients, tests and results, and treatments; these data usually come from special-purpose clinical trials. Although the data are more detailed and the quality can be controlled more closely, the quantity is usually insufficient to insure adequate statistical power across specific patient characteristics. To insure fair and accurate interpretation of the data, both of these data-driven approaches require methods to adjust for differing case mixes of different providers. Furthermore, to be useful, user-friendly interfaces for both types of data-driven approaches must be developed to insure that healthcare administrators, caregivers, and consumers will make use of them.

INFOSLEUTH

A goal of the InfoSleuth project at MCC is to exploit and synthesize new technologies into a unified system that retrieves and processes information in a dynamic network of information sources. Among other projects with similar goals are TSIMMIS [1] and the Information Manifold [2]. The InfoSleuth project extends earlier work on heterogeneous database integration. InfoSleuth models dynamic, web-based environments, in which there is no formal control over the registration of new information sources. InfoSleuth applications are developed without complete knowledge of the resources that will be available when they are run. InfoSleuth integrates new developments, such as agent technology, domain ontologies, information brokering and internet computing, in support of mediated interoperation of data and services in an open, dynamic environment.

This information infrastructure can be applied to create networks of healthcare information systems from collections of unique databases developed by hospitals, managed care plans, and physicians.

InfoSleuth is designed to provide interfaces for a wide range of distributed resources and tools. InfoSleuth software agents communicate the semantics of their conversations through common *ontologies*, or controlled vocabularies representing the schematic metadata of each domain [3]. An InfoSleuth application is a collection of agents written in Sun Microsystems' Java language for portability and compatibility with popular Web browsers. The agents communicate via Knowledge Query Manipulation Language (KQML) [4], using multiple content languages for expressing message semantics. In addition to Knowledge Interchange Format (KIF), agents also support Logical Data Language (LDL) [5] to communicate their reasoning about constraints.

Each agent has an assigned role in InfoSleuth. The broker agent serves the names of other agents based on constraints posed in the query message (One distinction from CORBA is the use of semantic constraints). The ontology agent serves the set of ontologies supported by the InfoSleuth application. The resource agent translates data stored in some external repository between the repository's access methods and semantic structures and the query language and domain ontology of InfoSleuth. The multi-resource query agent handles the decomposition and distribution of sub-queries to various resource agents and then recomposes the results. The task execution agent provides the necessary flexibility to control arbitrary patterns of information workflow based on a library of declarative task specifications. The user agent maintains a user's state, and provides the system interface that enables a user to communicate with the system independently of location.

Agents pose requests in terms of a specific ontology, called the "domain ontology of the application," that provides a semantic framework for information activities in the domain of the user's interest. Dynamic growth of agent communities is supported by means of semantic brokering, which allows agents to identify potential collaborators based on their advertised capabilities. Distribution of the agents reduces demands on the computation and storage power of a user's local machine, which need only support a Java-capable Web browser; this also means that access to resources that have registered with the broker is in-

dependent of the user's location; furthermore, the user need know nothing about the physical location or structural characteristics of any resource.

InfoSleuth in Healthcare

The InfoSleuth architecture as it might be deployed in a healthcare application is illustrated in Figure 1. Several resource agents have been configured to interact with the system by mapping the local schemas of the hospital information systems they represent to the global healthcare ontology and advertising the portions of the global ontology about which they can answer SQL queries. The user communicates with the system by means of an applet in the HAA toolset that exchanges queries and results with a user agent configured for that user. When the user issues a query, the user agent finds a task execution agent to whom to assign the query task. During interaction with the broker agent, the query is decomposed to be delivered only to those resource agents who have advertised knowledge of the elements of the ontology involved in the query (selective brokering increases the scalability of the system by reducing the number of queries that would return empty answers).

The healthcare ontology itself can be constructed with interaction with OOHVR, the New Jersey Institute of Technology's Object–Oriented healthcare vocabulary resource [6]. More complex tasks involving human participants can be orchestrated by invoking the METEOR system (under development by the University of Georgia) for transactional workflow support [9].

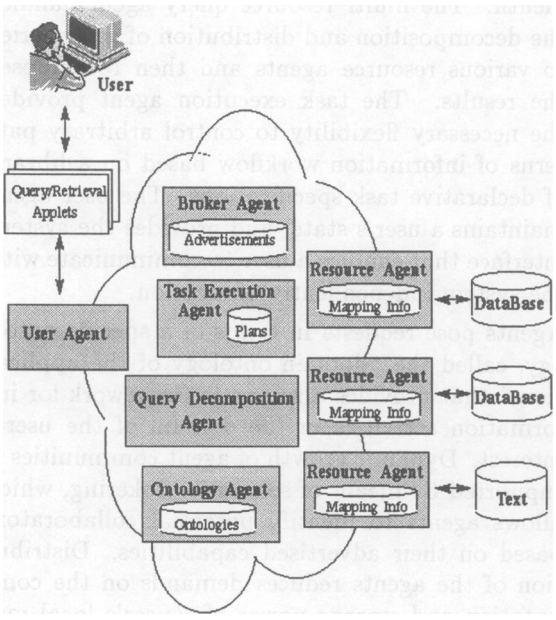

Figure 1 The InfoSleuth healthcare architecture

HEALTHCARE ADMINISTRATOR'S ASSOCIATE

Data mining is not as straightforward in healthcare as in some other domains, such as business: On-Line Analysis Programs (OLAP) perform best when there are only a few relevant concepts that are explicitly represented in the structure of the database. However, healthcare databases do not match this model. Concepts like quality and cost of care are not explicitly represented in the databases, and there are many relevant variables.

For this reason, we are developing a two-level approach to healthcare data mining. The first level supports relatively sophisticated analyses, such as association rule finding [7]. A healthcare data mining expert uses the first level tool to identify a subset of database variables relevant to a particular cluster of treatments and to map these variables onto a simplified, common model that holds across different treatment clusters. The healthcare administrator then uses the second level tool, the Healthcare Administrator's Associate, to query the model created by the first level tool. These queries enable healthcare administrators to compare the performance of their facilities to other facilities, evaluate trends over time, and explore the impact of policy decisions on healthcare costs and quality.

The conceptual model illustrated in Figure 2 is intended to reflect loosely the mental model that an administrative user brings to this application. Regardless of the particular domain under consideration, it is assumed that there will be a set of patient risk factors that determine which of a set of treatments a patient receives, and that these

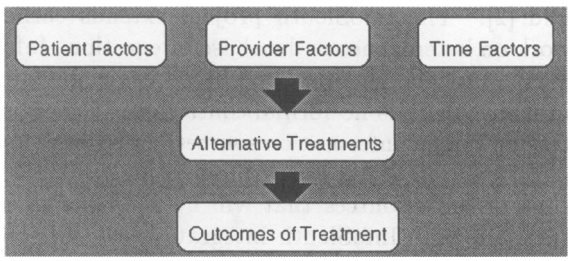

Figure 2 The structure underlying data mining analysis

patient risk factors and the applied treatment will determine the outcomes of treatment, along with other factors. The data mining expert uses the first level tool, in particular an association rule finding algorithm [7], to identify variables from the database that are statistically related to the choice of alternative treatments. The expert then uses domain knowledge to determine which of these variables constitute patient risk factors and which constitute outcomes. Patient risk factors can refer to database variables corresponding to primary or secondary diagnoses from a previous hospital encounter, or from the current encounter, when it can be inferred that these diagnoses were likely to have preceded the treatment. Patient risk factors can also include patient demographic data, such as age, and gender. Outcomes can refer to database variables that reflect the costs of care, such as aggregate charge data or length of stay; or the quality of care, such as diagnoses that reflect complications, disposition of the patient (for example, mortality), and any subsequent short-term readmittance to the hospital. Finally, the data mining expert specifies a set of healthcare provider factors, such as a specified set of hospitals, physicians, or policies; as well as time frames, as potential determining influences on patient risk factors, treatment, and outcomes.

The result of the data mining expert's efforts is a model of the variables, accessible through databases, which influence choice of treatment, outcomes of treatment, and patient risk factors. Unlike Bayesian networks [8], the current version of the system makes no assumptions about the statistical independence of these variables. Although it may be necessary to move toward a system with such simplifying assumptions, the present approach assumes that a very large amount of data will be available through InfoSleuth's access to multiple, heterogeneous databases. The statistical/causal relations between the variables are therefore directly calculated from the joint probability table of this set of variables.

The model and the accompanying databases constitute a knowledge base, or expert system, that can be queried by the second level tool, the Healthcare Administrator's Associate. This tool is implemented as a java applet that issues SQL queries to the InfoSleuth system to retrieve numerical data for calculating Bayesian probability values. Healthcare administrators issue their queries using a menu-driven interface to specify the factors of interest. The queries all take the form, "Tell me

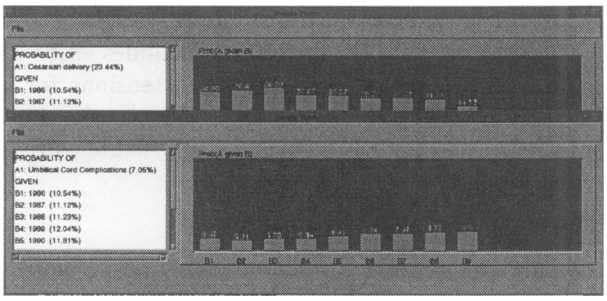

Figure 3 A prototype of HAA query results

the probability of X given Y," where X and Y are simple or complex predicates defined in terms of the identified variables, as listed in the applet's menus. A complex predicate corresponds to a set of simple predicates joined together by "OR's" and "AND's." The applet converts the Bayesian query into a set of SQL queries to the InfoSleuth system to count the frequencies of occurrence for various predicates. The probability values are calculated from these frequency counts, as follows:

$$\text{Prob}(X \mid Y) = \text{frequency}(X \cap Y)/\text{frequency}(Y)$$

The frequency of $X \cap Y$ corresponds to a count of the number of cases in which both X and Y predicates are true. The results are returned in the form of graphical bar charts, such as those depicted in Figure 3, that enable comparison of results across related factors.

DISCUSSION

We are beginning a process of evaluation that will feed back into the development process to improve the utility of the HAA. Much of this development will hinge on user interface. Future versions of the HAA will move computational responsibility from the applet to a specialized class of analysis agents. This will allow user interfaces to be developed independently of the analysis mechanisms.

A major issue in any distributed system is response time. The notoriously long running time of data mining processes is compounded by the effects of distribution. Although we have not yet performed significant optimizations of either system or code, we can nonetheless claim some benefit from our approach to distributed database access. Namely, the user time and effort involved in locating and extracting the appropriate data from several systems, and then combining the results meaningfully, would vastly outweigh any impact of distribution on response time of the current HAA.

Security of hospital information systems is a significant issue [10]. Our design includes embedded Java classes that can support extensions for authentication and encryption. The security model under development will support flexible privacy policies.

CONCLUSION

Much has been made in the popular and academic press of the comparison between the Internet and the United States' Interstate Highway System. This is a reasonable analogy; however, to say that, given the computer network interconnection provided by the Internet is the *entire* "information infrastructure" necessary to effect data communications would be to suggest by extension that the concrete and asphalt that comprise the physical highway system were useful without maps or roadsigns. On the contrary, maps came first.

By the same token, the Internet will become truly useful as a means of information sharing among geographically dispersed healthcare concerns when the roadmaps of healthcare, that is to say common structured vocabularies, serve to guide the development of the next generation of healthcare information systems, rather than traveling stoplight–infested twolanes between parochial homegrown data repositories.

Acknowledgments

This work was supported in part by the National Institute of Standards and Technology through HIIT Program Cooperative Agreement #70NANB5H1011.

InfoSleuth is an MCC consortial project sponsored by Texas Instruments, Computing Devices International, National Security Agency, Eastman Chemical and Hughes.

The InfoSleuth project is a large effort. A non-exhaustive list of contributors includes Roberto Bayardo, Bill Bohrer, Richard Brice, Andrzej Cichocki, Sumi Helal, Mike Huhns, Vipul Kashyap, Tomasz Ksiezyk, Marian Nodine, Brad Perry, Nancy Perry, Mosfeq Rashid, Marek Rusinkiewicz, Ray Shea, Munindar Singh, C. Unnikrishnan, Amy Unruh, and Darrell Woelk.

Our thanks to Ivan Shevchenko for his insights.

References

1. Garcia-Molina H., Hammer J., Ireland K., et al. Integrating and accessing heterogeneous information sources in tsimmis. In *Proceedings of AAAI Spring Symposium on Information Gathering*, 1995.

2. Levy A., Srivastava D., Kirk T. Data model and query evaluation in global information systems. *Journal of Intelligent Information Systems*, 5(2), Sept. 1995.

3. Bayardo R., Bohrer W., Brice R., et al. Infosleuth: Semantic integration of information in open and dynamic environments. In *Proceedings of the 1997 ACM International Conference on Management of Data (SIGMOD)*, pages 195–206, Tucson, Arizona, May 1997.

4. Finin T., Fritzson R., McKay D., McEntire R. KQML as an agent communication language. In *Third International Conference on Information and Knowledge Management*, Nov. 1994.

5. Zaniolo C. The logical data language (ldl): An integrated approach to logic and databases. Technical Report STP-LD-328-91, MCC, 1991.

6. Gu H., Cimino J. J., Halper M., Geller J., Perl Y. Utilizing OODB schema modeling for vocabulary management. In *Proceedings 1996 AMIA Annual Fall Symposium*, pages 274–278, Washington, DC, Oct. 1996.

7. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

8. Heckerman D. *Bayesian Networks for Knowledge Discovery*, chapter 11, pages 273–306. AAAI Press/ MIT Press, 1996.

9. Sheth, A., Kochut, K., Miller, J., et al. Supporting state-wide immunization tracking using multi-paradigm workflow technology. In *Proceedings of the 22nd Intl. Conf. on Very Large Databases*, September 1996.

10. Wiederhold G., Bilello M., Sarathy V., Qian X. L. A security mediator for health care information. In *Proceedings 1996 AMIA Annual Fall Symposium (Formerly SCAMC)*, pages 120–124, Washington, DC, Oct. 1996.